Semantic-Aware Contrastive Learning for Multi-object Medical Image Segmentation

Ho Hin Lee **Yucheng Tang** **Qi Yang** **Xin Yu** **Leon Y. Cai** **Lucas W. Remedios**

Shunxing Bao **Bennett A. Landman** **Yuankai Huo**

Department of Computer Science, Vanderbilt University
Department of Electrical and Computer Engineering, Vanderbilt University
Department of Biomedical Engineering, Vanderbilt University
Nashville, TN 37235
yuankai.huo@vanderbilt.edu

1 Network Architecture and Training Details

1.1 Network Architecture

We use a deeplabv3+ network structure as the segmentation backbone with ResNet-50 based encoder, consisting of 16 convolutional blocks in total, following with (each convolutional layer followed with 1 batch normalization layer and 1 activation layer (ReLU)):

- A convolutional layer with kernel size of 7×7 with 64 channels and stride size of 2 and a maxpooling layer with stride size of 2.
- 3 convolutional blocks with following 1×1 with 64 channels, 3×3 with 64 channels and 1×1 with 256 channels.
- 4 convolutional blocks with following 1×1 with 128 channels, 3×3 with 128 channels and 1×1 with 512 channels.
- 6 convolutional blocks with following 1×1 with 256 channels, 3×3 with 256 channels and 1×1 with 1024 channels.
- 3 convolutional blocks with following 1×1 with 512 channels, 3×3 with 512 channels and 1×1 with 2048 channels.

During pre-training stage (Stage 1 in Fig. 2), we apply a projection network p consisting of two dense layers with output dimension of 2048 and 256. After the encoder is well pre-pretrained with our proposed method, the small network p is discarded and the feature representation computed from the encoder is directly input into atrous spatial pyramid modules, consists of:

- 1 convolutional layers with kernel size of 1×1 with 512 channels
- 1 convolutional layers with kernel size of 3×3 with 512 channels, dilation rate of 6 and padding rate of 6
- 1 convolutional layers with kernel size of 3×3 with 512 channels, dilation rate of 12 and padding rate of 12
- 1 convolutional layers with kernel size of 3×3 with 512 channels, dilation rate of 18 and padding rate of 18
- 1 adaptive average pooling layer with size 1 and 1 convolutional layer with kernel size of 1×1 with 512 channels

Preprint. Under review.

The representation is separately input into each layer and the output from each layer is concatenated to input into a small decoder with two convolutions layers of 1×1 kernel size only with 512 channels and the total number of the semantic target classes respectively (No batch normalization and activation layer for the final convolutional output). The final output is bi-linearly upsampled to the same dimension of the input images.

1.2 Training Details

For training with medical imaging datasets, 5-fold cross-validation is performed for both contrast-enhanced phase and non-contrast phase CT (Training: 60 volumes (contrast-enhanced) and 44 volumes (non-contrast), validation: 20 volumes (contrast-enhanced) and 6 volumes (non-contrast), and testing: 20 volumes (contrast-enhanced) and 6 volumes (non-contrast)). For stage 1 training, we extract 30 2D query patches of each target in each axial slices to ensure that patches are fully covered the region of the corresponding organs with significant variation of anatomical morphology. More than 400k patches with dimensions 128×128 are used to train with stochastic gradient descent (SGD) optimizer for 5 epochs with a batch size of 4 and learning rate of 5×10^{-4}. We have evaluated the variation of the temperature parameter towards the segmentation performance and $\mathcal{T} = 0.1$ achieved the best performances across all other temperature values. For segmentation task, the encoder's weight is frozen and the decoder with ASPP module is trained for 10 epochs with Adam optimizer with batch size of 4 and learning rate of 10^{-4}. We used the validation set to choose the model with the highest mean Dice score for all semantic targets segmentation and perform testing evaluation as the quantitative representation on the testing set.

For training with natural imaging dataset, the augmented data with well annotations of [5] are used, resulting with 10582 images for training, 1449 images for validation and 1456 images for testing. A 2D segmentation model is initially trained with Deeplabv3+ network structure using ResNet-50 as the encoder backbone for 23 epochs. SGD optimizer with batch size of 8 and learning rate of 10^{-3} are used to optimize the training process for the coarse segmentation model. 2D query patches are extracted corresponding to the coarse attention map as same as with the medical imaging dataset. The extracted patches are then trained with the same network architecture with that with medical imaging dataset, but with $\mathcal{T} = 0.07$ and color distortions as addition to the data augmentation. For segmentation task, the encoder's weight is also frozen and the decoder with ASPP module is trained with Adam optimizer with batch size of 8 and learning rate of 10^{-4} for 10 epochs.

1.3 Details of Training Time

We train all models on NVIDIA 2080 Ti 11GB GPU with Pytorch implementation. The training time for medical imaging dataset is approximately: a) 10 hours for AGCL pre-training (per epoch) and b) 3 hours for training segmentation model (per epoch). For natural imaging dataset, it is approximately: a) 8 hours for coarse segmentation model (complete 23 epochs), b) 6 hours for AGCL pre-training (per epoch), and c) 2 hours for training segmentation model (per epoch).

1.4 Prepossessing

For medical imaging dataset, we apply the prepossessing steps as following: (i) applying soft tissue windowing within the range of -175 to 250 Hu and perform intensity normalization of each 3D volume, v with min-max normalization: $(v - v_1)/(v_{99} - v_1)$, where v_p denote as the p^{th} intensity percentile in v, and (ii) apply volume-wise cropping in z-axis with body part regression algorithm to extract the abdominal region only for segmentation and ensure the similar field of view between scans [10]. The coarse segmentation is computed with a coarse-to-fine pipeline [8] to ensure the robustness of organ localization and reduce the chance of extracting biased organ information for contrastive pre-training.

For natural imaging dataset, we follow [2, 3] to preprocess the input images to dimensions 513×513 for training the coarse segmentation model (with random horizontal flipping and scaling). Color distortion is further added into the data augmentation for AGCL pre-training stage.

Table 1: Ablation studies of non-contrast dataset on the segmentation performance in various network backbones.

Encoder	Pretrain	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stomach	Aorta	IVC	PSV	Pancreas	R.A
ResNet50	×	0.960	0.918	0.921	0.754	0.783	0.964	0.950	0.840	0.839	0.691	0.796	0.372
ResNet50	SimCLR	0.964	0.938	0.946	0.800	0.801	0.969	0.946	0.901	0.869	0.739	0.804	0.386
ResNet50	CE	0.972	0.952	0.961	0.812	0.859	0.974	0.959	0.934	0.914	0.768	0.838	0.551
ResNet50	AGCL	**0.982**	**0.962**	**0.965**	**0.834**	**0.879**	**0.982**	**0.967**	**0.945**	**0.929**	**0.790**	**0.850**	**0.560**
ResNet101	×	0.949	0.893	0.912	0.721	0.739	0.956	0.920	0.768	0.776	0.653	0.772	0.362
ResNet101	SimCLR	0.957	0.932	0.937	0.775	0.759	0.964	0.931	0.893	0.869	0.674	0.771	0.439
ResNet101	CE	0.970	0.943	0.955	0.783	0.811	0.972	0.943	0.911	0.897	0.722	0.811	0.501
ResNet101	AGCL	**0.977**	**0.956**	**0.965**	**0.811**	**0.870**	**0.979**	**0.966**	**0.922**	**0.915**	**0.799**	**0.842**	**0.573**

Table 2: Comparison of the fully-supervised, unsupervised, semi-supervised and partially supervised state-of-the-arts on the non-contrast testing dataset. Our method achieves the best Dice score, mean surface distance and Hausdorff distances. (We show 8 main organs Dice scores due to limited space, $*$: fully-supervised approach, $\star$: semi-supervised approach, $\triangle$: partially supervised approach.)

Method	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Aorta	IVC	Average Dice	Mean Surface Distance	Hausdorff Distance
$Cicek\ et\ al.$ $*$[4]	0.937	0.856	0.912	0.690	0.631	0.920	0.880	0.769	0.762	4.235	27.102
$Roth\ et\ al.$ $*$[9]	0.940	0.890	0.923	0.701	0.724	0.948	0.878	0.770	0.771	4.668	27.183
$Heinrich\ et\ al.$ [6]	0.910	0.865	0.889	0.624	0.656	0.930	0.860	0.759	0.748	4.976	30.241
$Zhu\ et\ al.$ $*$[13]	0.948	0.880	0.920	0.710	0.734	0.950	0.879	0.803	0.790	3.850	24.384
$Lee\ et\ al.$ $*$[8]	0.954	0.874	0.928	0.701	0.753	0.958	0.897	0.794	0.798	3.481	21.733
$Zhou\ et\ al.$ $\triangle$[12]	0.960	0.900	0.943	0.739	0.810	0.965	0.920	0.810	0.833	3.213	18.866
$Chaitanya\ et\ al.$ $\star$[1]	0.969	0.940	0.955	0.910	0.834	0.970	0.911	0.867	0.854	2.732	15.023
$Wang\ et\ al.$ $\star$[11]	0.979	0.961	0.964	0.941	0.865	0.979	0.937	0.923	0.887	2.210	12.419
$Khosla\ et\ al.$ $*$[7]	0.975	0.952	0.962	**0.943**	0.857	0.976	0.925	0.915	0.879	2.024	14.070
Ours (SSCL) $\star$	0.957	0.932	0.937	0.800	0.801	0.969	0.901	0.869	0.848	2.567	15.889
Ours (AGCL)$*$	**0.982**	**0.962**	**0.965**	0.834	**0.879**	**0.982**	**0.945**	**0.929**	**0.892**	**2.013**	**12.315**

References

[1] Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. arXiv preprint arXiv:2006.10511 (2020)

[2] Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)

[3] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)

[4] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)

[5] Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: 2011 International Conference on Computer Vision. pp. 991–998. IEEE (2011)

[6] Heinrich, M.P., Maier, O., Handels, H.: Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities. VISCERAL Challenge@ ISBI **1390**, 27 (2015)

[7] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020)

[8] Lee, H.H., Tang, Y., Bao, S., Abramson, R.G., Huo, Y., Landman, B.A.: Rap-net: Coarse-to-fine multi-organ segmentation with single random anatomical prior. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1491–1494. IEEE (2021)

[9] Roth, H.R., Shen, C., Oda, H., Sugino, T., Oda, M., Hayashi, Y., Misawa, K., Mori, K.: A multi-scale pyramid of 3d fully convolutional networks for abdominal multi-organ segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 417–425. Springer (2018)

[10] Tang, Y., Gao, R., Han, S., Chen, Y., Gao, D., Nath, V., Bermudez, C., Savona, M.R., Bao, S., Lyu, I., et al.: Body part regression with self-supervision. IEEE Transactions on Medical Imaging **40**(5), 1499–1507 (2021)

[11] Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3024–3033 (2021)

[12] Zhou, Y., Li, Z., Bai, S., Wang, C., Chen, X., Han, M., Fishman, E., Yuille, A.L.: Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)

[13] Zhu, Z., Xia, Y., Xie, L., Fishman, E.K., Yuille, A.L.: Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma. In: International conference on medical image computing and computer-assisted intervention. pp. 3–12. Springer (2019)